UNITED STATES PATENT APPLICATION

FOR

METHOD AND SYSTEM FOR SWITCHING VARIABLE SIZED PACKETS

Inventor:

Ofer Iny

Joseph A. Sawyer, Jr.
Sawyer Law Group LLP
2465 E. Bayshore Road, Suite 406
Palo Alto, California 94303

METHOD AND SYSTEM FOR SWITCHING
VARIABLE SIZED PACKETS

CROSS-REFERENCE TO RELATED APPLICATION

This application is claiming, under 35 USC 120, the benefit of provisional patent

application Serial No. 60/252,470 filed on November 22, 2000.

5   ## FIELD OF THE INVENTION

The present invention relates generally to a variable size data packet switching

device and more particularly to a system for switching variable size packets in a network.

BACKGROUND OF THE INVENTION

Modern data networks rely on a variable size packet transport network to

interconnect the various network elements. Packet switching devices are required to route a

packet through a network from a source to a destination. Typically a switching device has a

plurality of ports. Data packets arrive through one of the ports and are routed out one or a

plurality of ports.

15   A switching device, having a plurality of input and output ports, is required to

support transporting variable sized packets from inputs to outputs while maintaining packet

ordering within a flow. A flow is defined as a stream of packets arriving from one specific

source to one destination. It is desirable that a switching device be scalable such that more

inputs and outputs may be added, preferably while it is operating, while maintaining the

20   same performance properties.

A scalable switching device can be separated into three parts: an ingress controller,

an interconnect network, and an egress controller. Typically the ingress controller segments variable sized packets into fixed size cells. The cells are then routed through the interconnect network to the designated output. The egress controller then reassembles the cells into packets and reorders the packets to recover the ingress order.

5   A scalable interconnect network, referred to as a fabric, may be a multi-stage network where multiple paths exist from ingress to egress. In this case two categories of routing cells from input to output may be defined. Static Routing (SR) refers to a method where a path through the fabric is pre-determined for each flow. Dynamic Routing (DR) refers to a method where cells of a flow may take different paths. The advantage of SR is

10   that cells arrive at the output in order per flow. However, significant inefficiencies result from blocking, where one flow happens to select the same fabric link as another and by doing so oversubscribing the link capacity. Accordingly, dynamic routing (DR) is a preferred method for routing cells. DR greatly reduces the blocking problem. However, cells from a flow may arrive misordered and interleaved with cells from other flows.

          The problem of misordering may be divided into two parts. First, flow cell

15   reordering, and second whole packet reassembly and reordering. Typically, each problem was solved separately in dynamic routing fabrics.

          Accordingly, what is needed is a system which solves these problems differently. The present invention addresses such a need.

20

SUMMARY OF THE INVENTION

          A system for switching variable size packets in a network is disclosed. The system comprises at least one ingress controller which receives a plurality of packets and which

1964 P   -2-

segments each of the packets into fixed sized fragments. At least one ingress controller has a time-clock. All ingress controller's time-clocks are synchronized to within a tolerance. Each fragment is tagged with at least a unique source of ID, time-stamp, and a fragment-number to form a cell. Each cell belonging to one packet has the same time-stamp value. The ingress controller sends each of the cells through a link such that a cell's destination is reachable through that link. The system includes a fabric element which receives cells from a plurality of inputs links. The cells are ordered. The fabric element sends the ordered cells through a plurality of outputs and through which a cell's destination is reachable. The cell order is defined such that a cell ahead of another either has a lagging time stamp, or if the timestamp is the same the cell ahead of another has a source-id which has a predetermined priority, or if both the timestamp and the source-id are the same the cell ahead of another has a lagging fragment-number. The system finally includes at least one egress controller which receives the ordered cells from the plurality of input links, and sends the ordered cells through an output where such order results in complete packets.

A packet switching device in accordance with the present invention solves the cell ordering and packet reassembly issues using a unified distributed method in a multi-stage interconnect network.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a block diagram of a packet switching device.

Figure 2 is a block diagram of the ingress controller.

Figure 3 is the format of a data cell from the ingress controller to the fabric, and from fabric element to fabric element, and from the fabric to the egress controller.

Figure 4 is a block diagram of a multistage fabric.

Figure 4a is a block diagram of a multistage fabric plane implemented with a number of fabric elements.

Figure 5 is a block diagram of a fabric element.

Figure 6 is a block diagram of an egress controller.

DETAILED DESCRIPTION

The present invention relates generally to a variable size data packet switching device and more particularly to a system for switching variable size packets in a network. The following description is presented to enable one of ordinary skill in the art to make and use the invention and is provided in the context of a patent application and its requirements. Various modifications to the preferred embodiment and the generic principles and features described herein will be readily apparent to those skilled in the art. Thus, the present invention is not intended to be limited to the embodiment shown but is to be accorded the widest scope consistent with the principles and features described herein.

A method and system in accordance with the present invention, a multi-stage interconnect network (MIN), fabric, is built out of fabric elements connected in stages where each fabric element of a specific stage is connected to several fabric elements of the next stage. The MIN is used to connect ingress and egress controllers. The MIN has several routes from an ingress to egress. In a dynamic routing (DR) scheme, the ingress controller and the MIN routes cells to their indicted destination while attempting to balance the load on the available internal links. The Ingress controller constantly sends data cells on all output links. Data cells may have valid packet fragment, full, or may be empty. Other unrelated

cells may be interleaved among the data cells through the same links.

Variable sized packets entering through the ingress controller are segmented into fixed size fragments. The fragments are tagged with a destination, timestamp, unique source-id, and a fragment-number to form a data cell. Data cells from same packet have the same

5   timestamp. The ingress controller selects an output link for a cell such that the cell's indicated destination is reachable through the link while maintaining load balance over all possible links. When cells with packet fragments are not available for transmission on a link the ingress controller sends empty data cells, indicated by a cleared fragment valid flag, with the current timestamp, and unique source-id. Data cells on all output links are always

10   ordered.

Cell order is defined such that a cell ahead of another has a lagging timestamp, or if the timestamp is the same has a source-id which has a predetermined order, or, if both the timestamp and the source-ids are the same, has a lagging fragment-number. Cell output order is a sequence of ordered cells where all cells are destined to the output and all cells of each

15   packet destined to that output are present.

A fabric element (FE) has a FIFO per input link. An arriving data cell is buffered in its respective FIFO if the cell has a packet fragment, or if the FIFO occupancy is below a threshold and the cell is an empty data cell.

The fabric element sorts the oldest cells of all input FIFOs. The highest sorted cell is

20   selected if all active inputs FIFOs have at least one cell. An active link is one through which a data cell was received during a past period (empty or full). The FE has a FIFO per output link. If the selected cell has a packet fragment it is placed in one such FIFO. The output FIFO is selected such that the cell's indicated destination is reachable through the link while

maintaining load balance over all such links. When a data cell from an output FIFO is not available for transmission on a link the FE sends an empty data cell with the timestamp, source-id, and fragment-number of the last data cell that was selected from the sorter (full or empty). Thus, data cells on all output links are always ordered (with the exception of possible empty data cells with non-empty cells).

The egress controller has a FIFO per input link where arriving cells are buffered. The egress controller sorts the oldest cell in each FIFO. The highest sorted cell is selected for output if all active input FIFOs have at least one cell. If the selected data cell has a packet fragment it is placed in an outgoing buffer. As a result, cells in the output buffer are output ordered. That is, packets are fully reassembled and are ordered according to their chronological entry into the fabric.

To describe the present invention in more detail, refer now to the following description in conjunction with the following figures. Figure 1 is a block diagram of a packet switching device 10. Referring to Figure 1, the packet switching device 10 has a number of ingress controllers (ICs) 12, an interconnect network 14, and a number of egress controllers (ECs) 16. The ICs 12 and ECs 16 have a number of independent links to the interconnecting network 14 such that the external port capacity can be supported. In one implementation there are 32 such links from each of the ICs 12 to the interconnect network 14 and from the interconnect network 14 to each of the OCs 16.

Figure 2 is a block diagram of the ingress controller 12. Referring to Figure 2, the ingress controller 12 has an external packet interface 102, a packet segmenter 104, a global clock 106, a destination processor 108, and a fabric interface switch 112. Complete packets arrive through the packet interface 102. The packet segmenter 104 breaks the packets into

fragments, in one implementation this is a fixed 32-bytes fragment, and appends various tags to it to form a data cell. The segmenter 104 sends the data cells to the destination processor 108. The destination processor 108 sends each data cell to the fabric, through the fabric interface switch 112 and a fabric link, such that the cell's destination is reachable and all possible links are load balanced. One implementation has a reachability lookup table 110 where a cell's destination is looked up to get the possible output links. When there are no packet fragments, the ingress controller sends empty data cells with the timestamp set to equal the value of the global-time-clock 106 and source-id.

Figure 3 is a preferred embodiment of the format of a data cell 200 from the ingress controller to the fabric, and from fabric element to fabric element, and from the fabric to the egress controller. Referring to Figure 3, the cells from the ingress controller to the fabric elements have a Time Stamp 204, a Fragment-number 210, a Source-id 206, a Destination ID 208, and Fragment Valid (FV) flag 202. The FV flag 202 indicates if a packet fragment 212 is contained in the data cell. If the FV flag 202 is set, then the time stamp 204 is a copy of the global-time-clock in the ingress controller at approximately the moment the first cell of the packet (cell with the first data fragment) was sent to the fabric interface switch. Thus, each cell belonging to same packet has the same time stamp. If the FV flag is clear 202, then the time stamp is the value of global-time-clock when the empty cell was sent. The fragment-number 210 indicates the location of the cell in the packet. In one implementation it is an incrementing number starting at zero for the start of packet cell. The Source-id 206 is a unique global number. The destination-id 208 indicates the destination output port of the packet. The destination-id is irrelevant when the cell is empty (FV flag clear).

Cells sent from the Ingress controller are always ordered on any one link. That means

that a cell ahead of another one on a link has a lagging timestamp, or, if the timestamp is the same, a lagging fragment-number. In a pipelined implementation, the timestamp and fragment-number generation for both full and empty data cells must be consistent to result in this behavior.

Referring back to Figure 1, in one implementation the interconnecting network is made up of 32 independent fabric planes. A fabric plane can be comprised of one fabric element or of a number of fabric elements. In one implementation a fabric element has 64 or 32 independent inputs and 64 or 32 independent outputs.

Figure 4 is a block diagram of a multistage fabric. Referring to Figure 4, each plane of the interconnect network can be recursively built using a multi-stage network. An example of a known multistage network is shown where each fabric element 302a - 302n of the first stage is connected to all elements 304a – 304n of the second stage, and each element of the second stage is connected to each element 306a – 306n of the third stage.

Figure 4a is a block diagram of a multistage fabric plane implemented with a number of fabric elements. Referring to Figure 4a, the multi-stage fabric plane of figure 4 can be physically constructed out of fabric elements partitioned as shown. The first and third stage fabric elements are implemented in one fabric element device 402a – 402n and the second stage fabric element in another fabric element device 404a – 404n. Thus, figure 4a is a folded view of figure 4 along the center.

Figure 5 is a block diagram of a fabric element 500. Referring to Figure 5, the fabric element has a number of input interfaces through which it receives cells from the previous stage and a number of output interfaces through which it sends cells to the next stage.

The fabric element of size n x n has an input switch 502, n input FIFOs 504, sorter

506, destination processor 508, n output FIFOs 512, and output switch 514. Data cells arriving from the inputs through the input switch 502 are placed in the link's respective FIFO 504 if they contain a packet fragment or if the FIFO occupancy is below a threshold and they are empty cells. The cell sorter 506 reads the oldest cell from each input FIFO and

5 sorts the cells in order. Cell order is defined such that a cell ahead of another has: lagging time stamp, or if the timestamp is the same, has a source-id which has a predetermined priority (such as lower numerical value), or if both the timestamp and the source are the same has a lower fragment-number. When all incoming active link FIFOs have at least one cell (sorter has one cell from each FIFO) the sorter 506 removes the highest sorted cell. The

10 sorter 506 forwards that cell to the destination processor 508 if the cell has a data fragment. The sorter 506 remembers the timestamp, source-id, and fragment-number of the last removed cell.

The destination processor 508 examines the destination of the cell and selects one of the possible links through which the cell's destination is reachable while maintaining load balance over all possible links. It then places the cell in the selected output FIFO 512. Cells are sent from the output FIFOs 512 to the output links through the output switch 514. One implementation has a reachability lookup table 510 where a cell's destination is looked up to get the possible output links.

When no cells are available in an output link's FIFO then an empty data cell is sent

20 with the timestamp, source-id, and fragment-number of the last cell that was removed from the sorter. Cells sent from the Fabric Element are always ordered on any one link. Cell order is defined such that a cell ahead of another either has a lagging timestamp, or if the timestamp is the same, has a source-id which has a predetermined priority, or if both the

timestamp and the source-ids are the same, has a lagging fragment-number.

Figure 6 is a block diagram of an egress controller 16. Referring to Figure 6, an egress controller with n inputs has an input switch 602, n input FIFOs 604, sorter 606, a packet checker 608, and an output FIFO 610. Cells arriving from the inputs through the input switch are placed in their respective FIFO 604 if they are full or if the respective input FIFO is below a threshold and they are empty cells. The cell sorter 606 reads the oldest cell of each input FIFO and sorts the cells in order. Cell order is defined such that a cell ahead of another either: has: lagging a time stamp, or if the timestamp is the same, has a source-id which has a predetermined priority, or if both the timestamp and the source are the same, has a lagging fragment-number. When all incoming active links' FIFOs have at least one cell the sorter removes the top cell. If the cell contains a packet fragment then it is forwarded to the packet checker. The packet checker verifies that the cell is the expected one in the packet sequence and if so places it in the output FIFO. If the checker detects an incomplete packet that packet is deleted from the output FIFO.

Although the present invention has been described in accordance with the embodiments shown, one of ordinary skill in the art will readily recognize that there could be variations to the embodiments and those variations would be within the spirit and scope of the present invention. Accordingly, many modifications may be made by one of ordinary skill in the art without departing from the spirit and scope of the appended claims.